Data Science Projects
(D2K Learning Lab)

Statistics 435 / 535 • Fall 2018 • Genevera Allen

Time & Location:
Lecture: Wednesdays 4:00 - 5:30pm Location: Rayzor 113

Instructor:
Genevera Allen
Office: Duncan Hall 2080
Email: gallen@rice.edu
Office Hours: By appointment.

Teaching Assistant:
Tianyi Yao
Office: Duncan Hall 1041
Email: ty13@rice.edu
Office Hours: By appointment.

Course Description:
In this project-based course, student teams will complete semester-long data science research or analysis projects selected from a variety of disciplines and industries. Students will also learn best practices in data science.

Course Overview:
This is a project course where student teams will complete semester-long data science research or analysis projects selected from a variety of disciplines and industries. Each student team will have a data science faculty mentor as well as a discipline mentor. Student teams are required to meet weekly outside of class and meet at least twice a month with their mentors. During the semester, teams will work towards monthly project milestones and present their progress in two interim reports and oral presentations. At the end of the semester, each team will produce a final project report and give an oral presentation on their work in the D2K Learning Lab Showcase. Reports must adhere to computational and scientific reproducibility standards. This course will also have a lecture hour during which students will learn best practices along each step of the data analysis pipeline; these include data wrangling, exploratory analysis, modeling, and validation. Emphasis will be placed on best practices to achieve computational and scientific reproducibility of data analysis results.

Course Objectives & Outcomes:
In this course, students will engage in experiential learning where they learn best practices in data science by working on real-world data science projects. Students will learn how to wrangle, explore, and model their data as well as how to achieve a reproducible analysis pipeline and validate their data analysis findings. Through these projects, students will also learn team-work, leadership, project management, and communication skills. Students completing this course should be able to (i) successfully complete a data science research or analysis project, (ii) successfully devise and employ a computationally and scientifically reproducible data analysis pipeline, (iii) work collaboratively on a team to complete a data science project, write reports, and give oral presentations.

Stat 435 vs. 535:
Graduate students should enroll in Stat 535 and undergrads in Stat 435. Graduate students are expected to help lead their team as well as provide major technical contributions to the team project.

Coursework:
The major work of this course will be to conduct a data science research or analysis project and produce a team report in a computationally reproducible format (e.g. Rmd, Jupyter). Final reports will be due at the end of the semester with two interim progress reports. Teams will also give a short oral presentation accompanying each report. Further details and instructions on the reports and presentations as well as grading rubrics will be given in a separate document.

Grading:
Attendance: 10%
Team Participation: 15%
Oral Presentations: 15%
Report I: 10%
Report II: 10%
Final Report: 40%

Attendance Note: Students can miss one lecture and one scheduled team meeting without penalty. After that, there will be a 5% reduction in the attendance grade for each lecture or team meeting missed. No exceptions.

Team Participation Note: Team participation grades will be determined by peer-evaluation using the CATME system at three times throughout the semester. The last two CATME scores will be used to determine team participation grades. Specifically, CATME scores range from 5-25. Average scores from 18-25 are in the A-range, 14-17 B-range, 10-13 C-range, and so forth. This peer-evaluation system will be reviewed during the lecture hour.

Tentative Schedule:

August 22	Introduction & Meet Teams	
August 29	Data science pipeline + Teamwork	
September 5	Computational Reproducibility I	
September 12	Computational Reproducibility II	
September 19	Project Management + Communication	
September 26	Presentations I	Report I Due + CATME I Due
October 3	Presentations I contd.	
October 10	Validation I	Report I Revisions Due
October 17	Validation II	
October 24	Validation III	
October 31	Presentations II	Report II Due + CATME II Due
November 7	Presentations II contd.	
November 14	Discussion: Validation & Reproducibility	Report II Revisions Due
November 21	[Canceled before T-giving]	
November 28	**D2K Learning Lab Showcase**	
December 12		Final Report Due + CATME III Due

Students with a documented disability requiring accommodations should speak with the instructor during the first two weeks of class.

This syllabus is subject to change with reasonable advance notice by the instructor.